Intro & linear models (INLP ch. 1 & 2)

CS 685, Spring 2021

Advanced Topics in Natural Language Processing

http://brenocon.com/cs685

https://people.cs.umass.edu/~brenocon/cs685_s21/

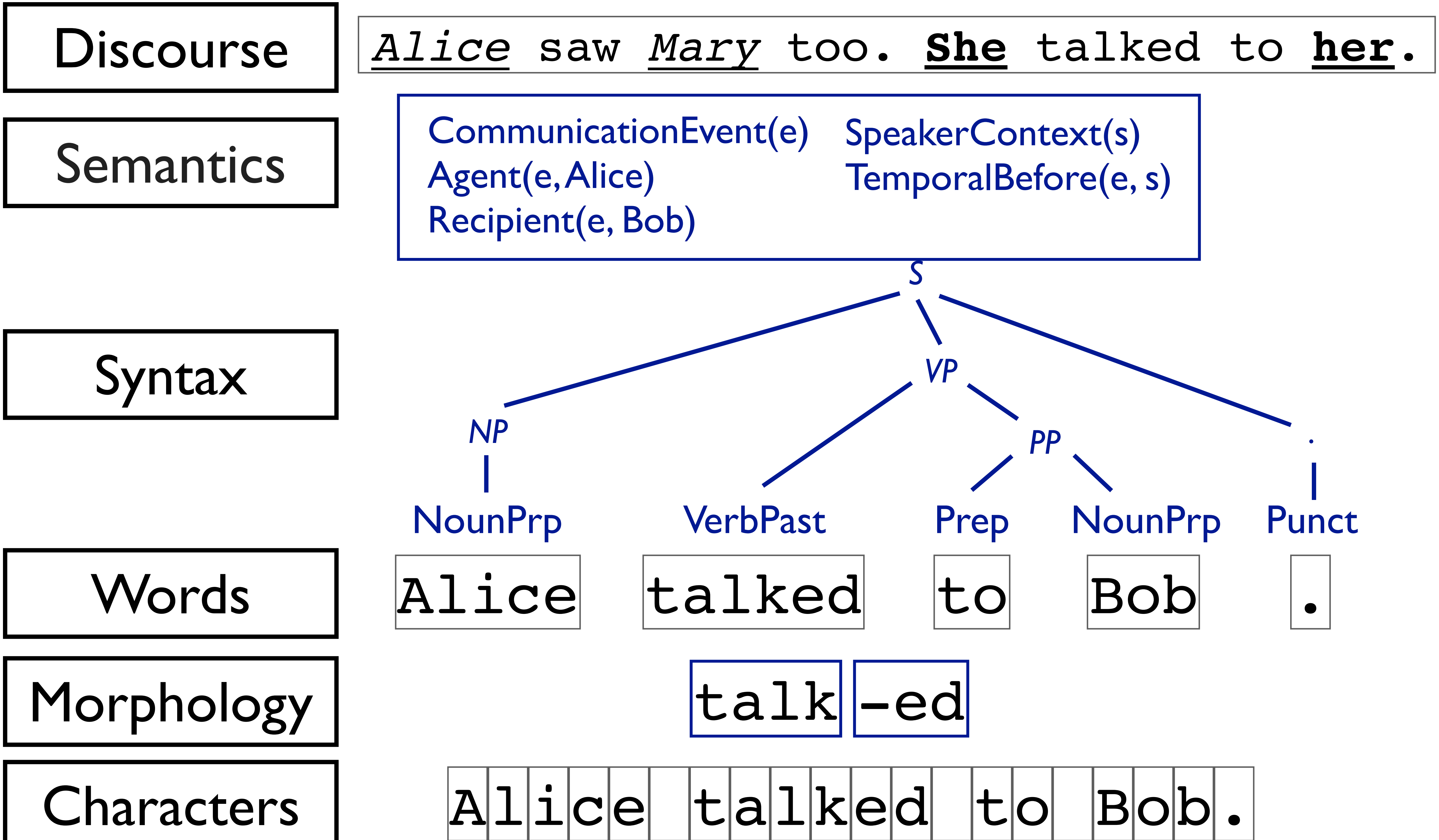
Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

[Material from Eisenstein (2019)]

Levels of linguistic structure



NLP as linguistic prediction

Scorer (model) Params

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \Psi(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

Predicted output

Possible outputs

Input (text)

Output candidate

- x: Text, y: Sentiment label
- x: Text, y: Syntax tree
- x: English text, y: Chinese text translation
- x: Book, y: Major characters & their relationships

NLP modeling

- **Pred./Search:** Scorer (model) Params

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \Psi(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

Predicted output

Possible outputs

Input (text)

Output candidate

- **Learning:** find a good θ from data
(we we need learning at all?)

- **Modeling:** design important ling. phenomena into Ψ
  - Reuse search/learning optimization methods for for many different NLP problems

NLP modeling

- **Pred./Search:** Scorer (model) Params

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} \Psi(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

Predicted output

Possible outputs

Input (text)

Output candidate

- Today: Linear models **Ψ**, BOW **x** & **f**, multiclass **y**
- Models: Naive Bayes and Logistic Regression
  - Learning: (regularized) MLE
- Wednesday: Neural network **Ψ**
- Later: sequential **x**
- Later: structured output **y**

Linear classification models

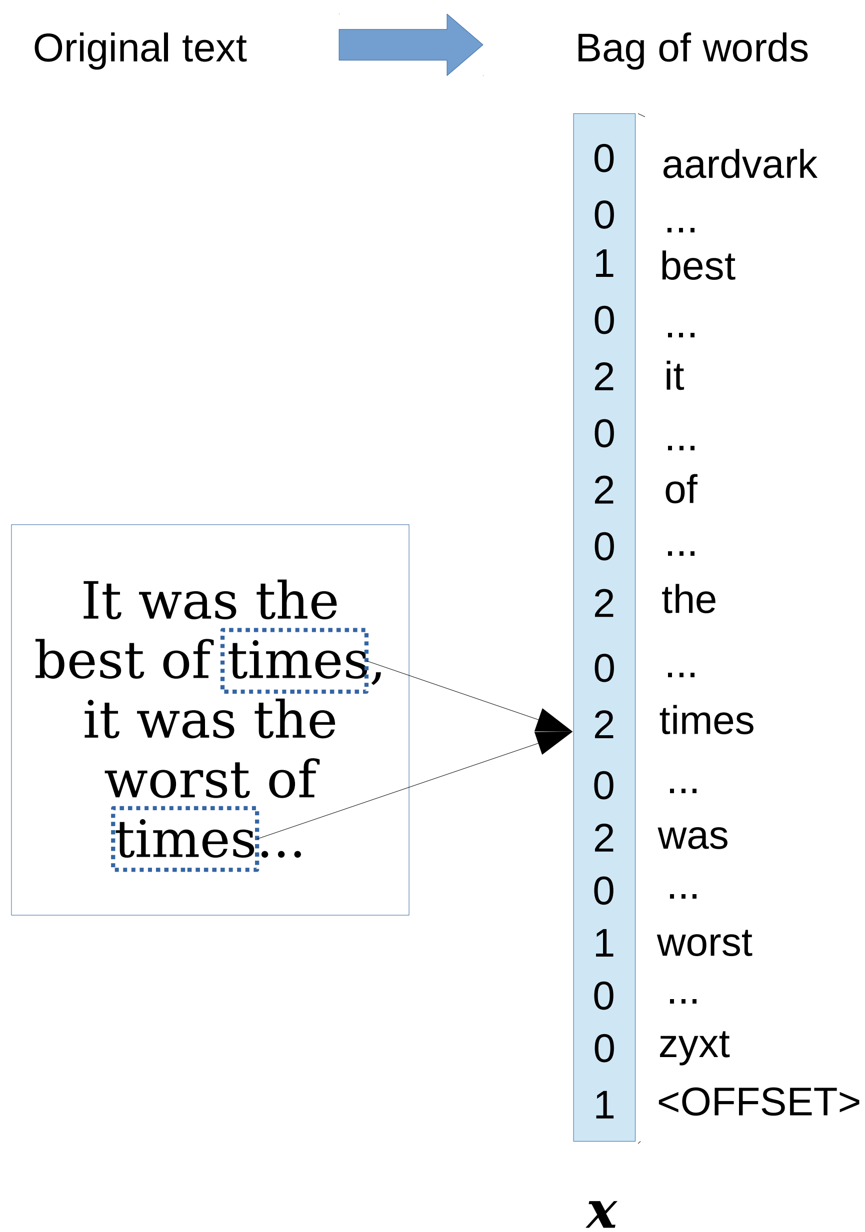
- Assume classification problem:
  - Input text **x**
  - Output discrete y**ϵY**, |**Y**|=K
- Scoring function is a dot product of weight vector θ and a vector-valued feature function **f**
  - **x** is a representation of the text. What to use?
  - **f** computes features to score the candidate output. What features to use?

Input (text)

Output candidate

$$\Psi(\boldsymbol{x}, y) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y) = \sum_j \theta_j f_j(\boldsymbol{x}, y)$$

Bag of words representation



- **x** is a vector, representing counts of words in the text
- Vocabulary **V**: set of all word types under consideration
  - Vocab size V = |**V**|
  - Len(x) = V
- BOW ignores order information!! Yet is useful….
- Idea: each word can a weights for each possible output category

$$\Psi(\boldsymbol{x}, y) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y) = \sum_j \theta_j f_j(\boldsymbol{x}, y)$$

Bag of words: linear model

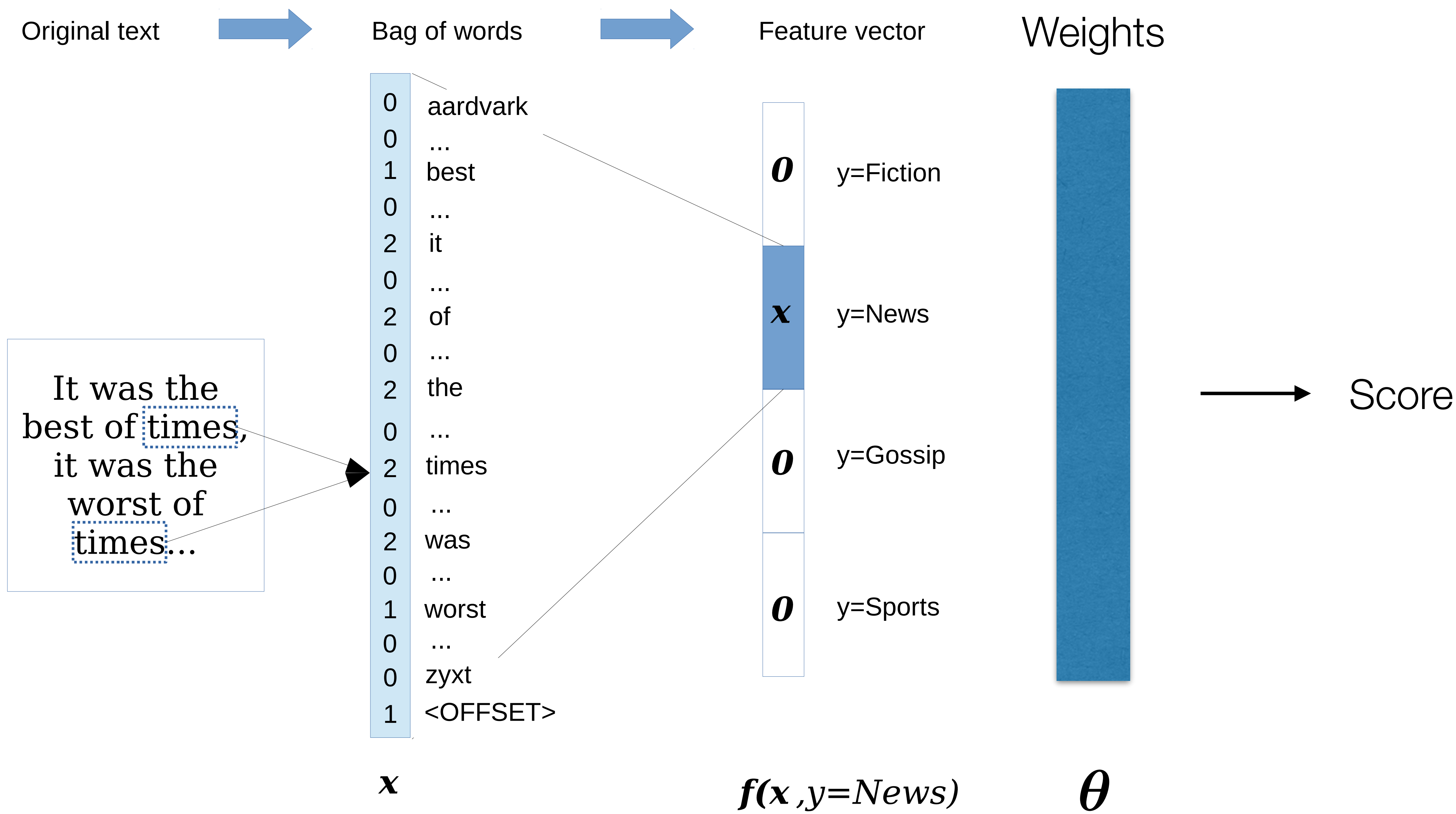
- One feature for each word and class pair.

$$f_j(\boldsymbol{x}, y) = \begin{cases} x_{whale}, & \text{if } y = \textsc{Fiction} \\ 0, & \text{otherwise} \end{cases}$$

- And another for y=NEWS, y=GOSSIP, y=SPORTS (and all other output classes)
- Thus
  - **f**: **X** x **Y** -> $\mathbf{R}^{VK}$
  - θ ϵ $\mathbf{R}^{VK}$

$$\Psi(\boldsymbol{x}, y) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y) = \sum_j \theta_j f_j(\boldsymbol{x}, y)$$

Bag of words: linear model



$$\Psi(\boldsymbol{x}, y) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y) = \sum_j \theta_j f_j(\boldsymbol{x}, y)$$

How to set parameters?

- Could use deterministic 1 and 0 weights — implicit in lexicon/dictionary/keyword methods (e.g. racial slur blacklist, sentiment lexicons, etc.)
- But if you have labeled data, typically **supervised learning** is better.
- Labeled data: "gold-standard" (text, label) pairs

$$\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$$

- Two linear, probabilistic models today
    - Naive Bayes
    - Logistic Regression

Naive Bayes

- Assume a model of both text and labels (each doc i.i.d.)

$$\mathrm{p}(\boldsymbol{x}^{(1:N)}, y^{(1:N)}) = \prod_{i=1}^{N} \mathrm{p}_{X,Y}(\boldsymbol{x}^{(i)}, y^{(i)})$$

- p(x,y) is a **generative model**: has a story of how both the label and document text was generated
  - generating text is a.k.a. **language model** — other LMs are used a lot in NLP
- Once we have a model, we can do:
  1. Learning: fit p(x,y)'s parameters to training data (using MLE)
  2. Prediction ("Search"): infer labels on new documents (using Bayes Rule)

NB generative model

Algorithm 1 Generative process for the Naïve Bayes classification model

for Instance $i \in \{1, 2, \ldots, N\}$ **do**:
  Draw the label $y^{(i)} \sim \text{Categorical}(\boldsymbol{\mu})$;
  Draw the word counts $\boldsymbol{x}^{(i)} \mid y^{(i)} \sim \text{Multinomial}(\boldsymbol{\phi}_{y^{(i)}})$.

Algorithm 2 Alternative generative process for the Naïve Bayes classification model

for Instance $i \in \{1, 2, \ldots, N\}$ **do**:
  Draw the label $y^{(i)} \sim \text{Categorical}(\boldsymbol{\mu})$;
  for Token $m \in \{1, 2, \ldots, M_i\}$ **do**:
    Draw the token $w_m^{(i)} \mid y^{(i)} \sim \text{Categorical}(\boldsymbol{\phi}_{y^{(i)}})$.

- Types vs. Tokens
- Generative probability notation:
  - a ~ Distrib(theta): "Random variable a is sampled according to distribution Distrib, parameterized by theta."
- Parameters
  - μ_k: prior probability of class k
  - $\phi_{k,w}$: probability word w gets generated under doc class k
- "Naive": each word token is generated independently.

NB prediction

- First assume we have parameters. How do we predict the label given text? Chose the one with *highest posterior probability* p(y | **x**) = p(**x**, y) / p(**x**)

$$\begin{aligned}\hat{y} &= \underset{y}{\operatorname{argmax}} \log \mathrm{p}(\boldsymbol{x}, y; \boldsymbol{\mu}, \boldsymbol{\phi}) \\ &= \underset{y}{\operatorname{argmax}} \log \mathrm{p}(\boldsymbol{x} \mid y; \boldsymbol{\phi}) + \log \mathrm{p}(y; \boldsymbol{\mu})\end{aligned}$$

$$\begin{aligned}\log \mathrm{p}(\boldsymbol{x} \mid y; \boldsymbol{\phi}) + \log \mathrm{p}(y; \boldsymbol{\mu}) &= \log \left[B(\boldsymbol{x}) \prod_{j=1}^{V} \boldsymbol{\phi}_{y,j}^{x_j} \right] + \log \mu_y \\ &= \log B(\boldsymbol{x}) + \sum_{j=1}^{V} x_j \log \phi_{y,j} + \log \mu_y\end{aligned}$$

- This can be shown to be a linear model (see text).
    - Parameters = log probs of words and class priors
    - Features = count of word under candidate class; candidate class

NB learning

- Intuitively, **relative frequency estimation** sounds good:

$$\phi_{y,j} = \frac{\text{count}(y, j)}{\sum_{j'=1}^{V} \text{count}(y, j')} = \frac{\sum_{i:y^{(i)}=y} x_j^{(i)}}{\sum_{j'=1}^{V} \sum_{i:y^{(i)}=y} x_{j'}^{(i)}}$$

- This has a deeper theoretical basis!

NB learning: MLE

- **Maximum Likelihood Estimation**: choose params that give highest likelihood to the training data

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathrm{p}(\boldsymbol{x}^{(1:N)}, y^{(1:N)}; \boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} \mathrm{p}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underbrace{\sum_{i=1}^{N} \log \mathrm{p}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta})}_{\mathcal{L}(\boldsymbol{\theta})}
\end{aligned}
$$

log likelihood function

NB learning: MLE

- Choose **ϕ,μ** to maximize log-likelihood

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \log \mathrm{p}_{\text{mult}}(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}_{y^{(i)}}) + \log \mathrm{p}_{\text{cat}}(y^{(i)}; \boldsymbol{\mu})$$

- Under the sum-to-1 constraints

$$\sum_{j}^{V} \phi_{y,j} = 1 \quad \forall y \qquad\qquad \sum_{k}^{K} \mu_k = 1$$

- Calculus with Lagrange multipliers
  ===> intuitive relative frequency estimates!

Bias-variance tradeoffs

- Does MLE overfit or underfit?
- Laplace smoothing: add a pseudocount,

$$\phi_{y,j} = \frac{\alpha + \mathbf{count}(y,j)}{V\alpha + \sum_{j'=1}^{V} \mathbf{count}(y,j')}$$

- (Why Vα ?)
- α=>large ==> ?

What about class priors?

- Choose **ϕ,μ** to maximize log-likelihood

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \log \mathrm{p}_{\mathrm{mult}}(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}_{y^{(i)}}) + \log \mathrm{p}_{\mathrm{cat}}(y^{(i)}; \boldsymbol{\mu})$$

- **μ** is set to class frequencies in training data. Is this realistic?
    - [See our paper! Keith and O'Connor, 2018]

Cond. indep. is a problem

- We can do better than BOW features by **feature engineering** lots of little variants of words and phrases
  - e.g. ngram features … character n-grams … words with or without lowercasing … number of digits in the text … number of punctuation marks in the text … etc.
  - Even overlapping features can have useful predictive value
- But… does NB do well with repetitive features?

$$\Psi(\boldsymbol{x}, y) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y) = \sum_j \theta_j f_j(\boldsymbol{x}, y)$$

Discriminative learning

- NB (generative) learning chooses params to maximize $p(X^{train}, Y^{train})$, then indirectly gives the linear prediction model.
- But if we just care about prediction, why not directly learn to minimize prediction errors?
  - Perceptron: choose params to minimize **1-0 loss**.
  - SVM: choose it to minimize **hinge loss**
- Logistic regression: choose theta to maximize **conditional log-likelihood** (a.k.a. minimize logistic loss a.k.a. cross-entropy)

Logistic regression

- Directly define the conditional probability of label given text via the **softmax** of the linear scoring function

$$\mathrm{p}(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y')\right)}$$

exp and normalize

- Learning: choose theta to maximize the **conditional log-likelihood**,

$$\log \mathrm{p}(\boldsymbol{y}^{(1:N)} \mid \boldsymbol{x}^{(1:N)}; \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \mathrm{p}(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

$$= \sum_{i=1}^{N} \underbrace{\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)})}_{\text{Give high scores to observed y(i)}} - \log \sum_{y' \in \mathcal{Y}} \exp\Big(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')\Big)$$

Logistic loss

Negative log-likelihood for one example

$$\ell_{\textsc{LogReg}}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) = -\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \log \sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y'))$$

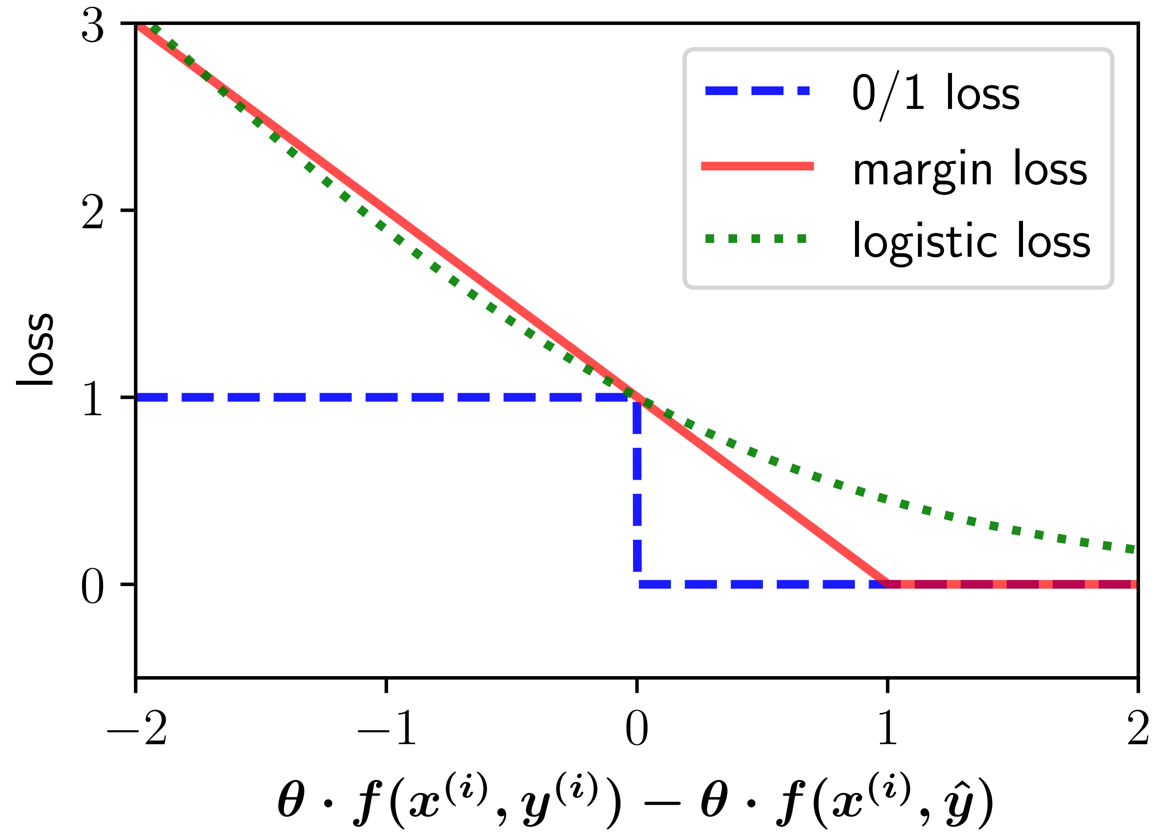


Figure 2.2: Margin, zero-one, and logistic loss functions.

- “Soft” grading of errors
  - 99% prob for y(i) => 😀
  - 1% prob for y(i) => 😂
  - 0%?

LogReg learning

- There is no closed form MLE
- But, fortunately, the log-lik is concave (NLL convex)
- Use gradient descent!
    - 1. Calculate gradient equations
    - 2. Use a batch or online gradient algorithm

Regularization

- For many NLP feature functions, training data is often linearly separable. Weights diverge to +/- inf
- Regularization is essential. Typically use the L2 norm of weights, resulting in a regularized loss:

$$L_{\textsc{LogReg}} = \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 - \sum_{i=1}^{N}\left(\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \log\sum_{y'\in\mathcal{Y}}\exp\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{x}^{(i)}, y')\right)$$

Summary: NLP prediction

Scorer (model) Params

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})}{\operatorname{argmax}} \Psi(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

Predicted output

Possible outputs

Input (text)

Output candidate

- Today: Linear models **Ψ**, BOW **x** & **f**, multiclass **y**
- Models: Naive Bayes and Logistic Regression
  - Learning: (regularized) MLE
- Wednesday: Neural network **Ψ**
- Later: sequential **x**
- Later: structured output **y**